Contents

Chapter 1

Introduction

The past decade social media has become ever more important in our daily lives. Nowadays most of us have, beside a real world life, a virtual life on social media. This increased engagement on social media may have a huge impact on the formation of opinions, not only on the individual scale, but also on a broader, societal scale. It is known that individuals not only form their opinions through self-reflection, but also through interactions with other people and with their surroundings [35]. The broad range of interactions people undergo on social media with people from all over the world can thus not be underestimated in the process of opinion formation. The understanding of the role of social media on the emergence of polarization and extremism in society is of uttermost importance.

One important feature of opinion dynamics on social media is the use of algorithmic personalization. Social media allow for simultaneous or asynchronous interactions between people without geographical constraints and thus allow for the spread of information at a faster pace and at an unprecedented scale [35]. People are, however, still bound by temporal and cognitive constraints. In order to assure a more pleasant/convenient experience on their platforms, social media companies use algorithms to order/filter the posts on an individuals time line according to what might be relevant to that individual [35]. It is thus important to know the possible effects of these filtering algorithms on the evolution of opinions in the population.

It is also known that some people change their opinions more easily than others, that is, once in a while you encounter somebody that is really stubborn and persistent towards its own opinion. One of the aims of this thesis is to design toy models of opinion dynamics with stubborn actors on theoretical and real-world social networks to analyze the interplay between individual resistance to change and the network structure in the evolution of two competing opinions.

1

Opinion dynamics models have two important layers. On one hand, we introduce social networks to describe the underlying structure of social interactions, while on the other hand we need appropriate opinion dynamics/formation models to reproduce the opinion dynamics and opinion formation processes found in real life.

1.1 Complex networks

Complex networks have become more and more popular for analyzing complex, dynamical systems. They are used in fields such as physics, economics, social sciences, biology, etc [12]. These different fields are very diverge, but they have at least one common ground: they often deal with a large number of variables/-components that interact with each other. In other words, in all these fields one encounters complex systems, systems where it is not possible to predict collective behavior based on the properties of the individual components alone [14]. Complex systems often display phenomena such as non-linearity, emergence, spontaneous order,... One of the tools to deal with these complex systems and to give us more insight in the possible underlying structures are complex networks. A network, also often called a graph, is a structure composed of nodes or vertices and a set of links or edges that indicate the interactions between the nodes [14]. Representing/modeling a complex system as a graph makes the system appear more simple and tractable, while it still includes the non-linearity of these systems. One can find the language to describe networks in mathematical graph theory. However, complex systems in real life situations often deal with a huge number of components, so the use of statistical and high-performance computing tools is inevitable [14]. One could argue that the study of complex networks lies somewhere at the intersection of mathematical graph theory and statistical physics [13].

The advantage of working with network models is that they reduce the level of complexity encountered in the real world, so that one can treat these systems in a more practical way. However we do want our models to display properties similar to the ones seen in real systems [14]. Since many real systems are not static but evolve in time, this means that we don't only need to deal with static networks but also with temporal networks. Temporal networks are networks where the edges are not always active, but instead become active for some periods of time [21]. In temporal networks time becomes an explicit element in the network representation [21]. Static networks have been widely studied and are often convenient for their analytic tractability, whereas temporal networks are, in some cases, more realistic [14].

Some other important concepts in network theory are the degree distribution, clustering/community structure, connectivity, etc. These concepts will be explained in depth in Chapter 2, Section 2.1, but let us already anticipate a bit on the case of social networks. It is found that many real-life social systems have a power law degree distribution [30]. This heterogeneous or scale free degree distribution represents one of the three general properties of social networks. The other two are short distances and high clustering, these two are also referred to as "small world phenomenon" (note that in the literature the term "small world" sometimes only refers to the property of short distances, however, in this thesis, we will use it to indicate both a high clustering and short distances) [30]. Ideally, our theoretical network should exhibit these three properties. Some theoretical networks, that are thoroughly studied, do not possess all of these three properties. However, they might still be used, because of their simplicity and ability to produce analytic results. When using these models one must always bear in mind their limitations to reproduce some properties encountered in real social systems.

One of the goals of this thesis is to investigate/determine the role of the underlying network structure on the formation and dynamics of opinions in the system. Some examples of theoretical network models used in this thesis are the Erdős-Rényi network, the stochastic block model, the Watts-Strogatz model, etc. These will be explained thoroughly in Chapter 2, Section 2.2.

1.2 Opinion dynamics

As stated before, the formation of an individual's opinion is the result of the interplay of many factors such as self-reflection, peer pressure, the individual's personality (eg. stubbornness), the information someone is exposed to, etc. Opinion dynamics models should try to capture these complex processes in a simplified way. Several models have been developed ranging from simple binary models to more complex, continuous approaches [37]. The basic idea of all opinion dynamic models is that the nodes or agents in a social network have a variable that represents their opinion and that is updated according to some predefined rules. These models are obviously a simplification of real world opinion dynamics. It is however shown that they do display a lot of aspects of real opinion formation such as agreement, transitions between order (consensus) and disorder (fragmentation), polarization, formation of echo chambers (clusters of people with the same opinion), etc [37].

The opinion of the agents in the model can be either discrete or continuous. This thesis deals with models where each agent can have one of the two opinions A or

B and thus only deals with the case of discrete opinions. This might come across as a huge simplification of the real world complexity of opinion formation, but also in real life people often have to choose between two competing opinions (eg. republicans or democrats, cat or dog, renting or buying a house,...).

The rules that determine how an agent updates his or her opinion can include many different aspects. They can, for example, be as simple as a majority rule (i.e. the majority model: if the majority of your neighbors have a certain opinion, you adopt that opinion as well) or can include a more probabilistic way of updating. One can also include a resistance parameter that determines the hesitation of an agent to change to a new opinion and/or a parameter that determines the influence of an agent on others. Furthermore, the agents can have a parameter that determines whether they are active or not.

Since this thesis deals with the particular case of opinion dynamics on on-line social platforms, it is important to define/set up different filtering algorithms. The filtering algorithms that real social media companies use, are corporate secrets, but we know that there are three main principles of content curation: popularity, semantic and collaborative filtering [35]. Here popularity filtering refers to the practice of promoting content that is popular across the platform; semantic filtering means that post similar to previous consumed posts are recommended and collaborative filtering suggests posts that are similar to the ones our friends consume [35]. In Chapter 3 a detailed description of the filtering algorithms used in this thesis will be given.

1.3   Statistical and social physics

Another important tool in the study of opinion dynamics/formations in large groups of people is the use of statistical physics. This branch of physics offers a framework that relates microscopic properties of atoms, molecules,... to macroscopic observed behavior and has a wide field of applications inside and outside the world of physics.

The observation that large number of people display collective, 'macroscopic' behavior begged for the use of the concepts and insights developed in statistical physics [37]. The application of the theory of statistical physics on social phenomena is referred to as social physics or sociophysics. The 'microscopic' constituents are now individual humans who interact with a limited number of other individuals and in that way form complex, 'macroscopic' groups such as human societies. These 'macroscopic' groups display stunning regularities, transitions from disorder to order, the emergence of consensus, universality, etc. The statistical physics approach of these social systems tries to explain the found regularities at large

scale as collective effects of the interactions among these individuals [37].

1.4 This thesis: hypotheses and goals

As said before, this thesis will try to investigate the interplay between the social network structure and the individual resistance to change in the evolution of two competing opinions. So on one hand, we want to determine the influence of different network structures on the formation and evolution of opinions, whereas, on the other hand, we would like to investigate the effect of stubborn agents in the network. Since this thesis deals with opinion dynamics on social media, the effect of filtering algorithms must also be included.

In one of their previous papers (see reference [35]) prof. Nicola Perra and prof. Luis E.C. Rocha investigated both the effect of different filtering algorithms and different network topologies on the formation and evolution of two competing opinions. They also considered the effect of nudging (meaning that one opinion is pushed to all agents in the network).
Their main findings were that the algorithmic filtering could not break the status quo when the prevalence of opinions was equally distributed in the population and that topological correlations (such as high clustering coefficient) could result in the formation of echo chambers [35]. On the other hand, a more heterogeneous contact pattern could hamper this formation of echo chambers [35]. If the two opinions were, initially, not equally distributed, one of the filtering algorithms (namely semantic filtering) caused a further increase in the predominant opinion [35]. In the case of nudging, they found that the population opinion moved to the nudged opinion relatively fast, even in the case of small nudging.
They, however, did not investigate the effect of stubborn agents in the network, nor did they investigate networks with a community structure. This thesis will build on their model and incorporate these new features. We will compare networks that have a community structure and relatively low clustering to networks without community structure, but with a high clustering coefficient (a detailed description of concepts like clustering coefficient and community structure will be given in Chapter 2, Section 2.1). We want to investigate whether networks with community structure are able to form echo chambers, such as one can observe in networks with a high clustering coefficient and, if so, if one can observe convergence to one opinion inside the communities. On the other hand, we want to determine the impact of stubborn agents on the formation and evolution of opinions in the network.

Chapter 2

Networks

2.1 Definition and network measurements

Networks or graphs are conceptually very simple and flexible objects that are made of a set of nodes/vertices V and a set of edges/links E, where the elements of E determine connections between elements of V [11]

$$E \subseteq \{\{x, y\} | x, y \in V\} \tag{2.1}$$

If self-loops are not allowed, we need the following condition on the elements of E

$$E \subseteq \{\{x, y\} | x, y \in V \text{ and } x \neq y\} \tag{2.2}$$

A graph with no self-loops is a simple graph. The nodes x and y of an edge $\{x, y\}$ are called the endpoints of the edge. It is possible that nodes are not joined by any edge, such nodes are called isolated or disconnected. If two or more edges have the same pair of endpoints, the graph is called a multi-graph. This thesis is not concerned with multi-graphs, nor are self-loops allowed. Graphs in which the edges have an orientation are called directed graphs. If the edges have weights w_{ij}, one speaks of a weighted graph.
A graph can be represented in different ways. One of the most common ways is by use of the adjacency matrix.

2.1.1   Adjacency matrix

The adjacency matrix **A** is a $N \times N$ matrix (for a graph with N nodes) where the elements indicate whether two nodes are connected by an edge or not [14]

$$A_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

For a simple, undirected graph, the adjacency matrix is symmetric ($A_{ij} = A_{ji}$) with zeros on the diagonal ($A_{ii} = 0$). Directed graphs have asymmetric adjacency matrices. If we are dealing with weighted graphs, the elements of the adjacency matrix are the weights w_{ij} of the edges

$$A_{ij} = \begin{cases} w_{ij} & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

with, generally, $0 \leqslant w_{ij} \leqslant 1$ [14].

2.1.2   Degree and degree distribution

The adjacency matrix contains a lot of information such as the degree of a node. The degree k_i of a node i is the number of edges attached to the node i or, in others words, k_i represents the number of nearest neighbors of the node i. The degree can be obtained from the adjacency matrix in the following way [14]

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{2.5}$$

In the case of a directed graph one can differentiate between the number of incoming edges k_i^{in} and the number of outgoing edges k_i^{out} of the node i. These are called the in-degree and out-degree respectively and are defined as [14]

$$k_i^{\text{in}} = \sum_{j=1}^{N} A_{ji} \quad \text{and} \quad k_i^{\text{out}} = \sum_{j=1}^{N} A_{ij} \tag{2.6}$$

The total degree of node i is then $k_i = k_i^{\text{in}} + k_i^{\text{out}}$. For weighted graphs the degree is easily generalized to the weighted degree s_i, often called strength [23]

$$s_i = \sum_{j=1}^{N} w_{ij} \tag{2.7}$$

From the degree of the nodes in the network one can construct the degree distribution. The degree distribution $P(k)$ represents the fraction of nodes with a degree k or, in other words, the degree distribution gives the probability that a randomly selected node has a degree k. The average degree can be obtained from the degree distribution [14]

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \sum_{k} k P(k) \tag{2.8}$$

where N is the total number of nodes in the network. The degree distribution also allows us to classify networks. The two most important classes are homogeneous and heterogeneous networks. Homogeneous networks have a bell curved degree distribution, eg. a Poisson distribution. In this case, most of the nodes have a degree closed to the average degree $\langle k \rangle$. Heterogeneous networks, on the other hand, have a power-law degree distribution, $P(k) \sim k^{-\gamma}$. They are also referred to as scale free networks, since they do not posses a characteristic length scale or, in this case, the average degree isn't a characteristic scale for the network [14].

2.1.3 Connectivity

In physics literature, the degree of a node is often called the connectivity [13][15]. This is different from the mathematical/graph definition of connectivity, where connectivity refers to the minimum number of nodes or edges that need to be removed to separate the remaining nodes in isolated subgraphs.
Two nodes u and v of a graph are connected if there exist a path between them. If the two nodes are connected by a path that contains only one edge, the two nodes are called adjacent.
The average degree $\langle k \rangle$ is a global measurement of the connectivity of the network [12]. If the average connectivity $\langle k \rangle$ is too small (this is, if there are too few edges), there will be many isolated nodes and only a few clusters with a small number of nodes. If more and more edges are added to the network, the small clusters will grow and will tend to connect to each other to form larger clusters. At some critical value of the connectivity, most nodes will be connected into a giant cluster (the giant component), which characterizes the percolation of the network [13].

2.1.4 Assortative and disassortative networks

Another concept that is closely related to the degree of a node is the notion of assortative and disassortative networks. More precisely, the notion of assortativity and disassortativity has to do with degree correlations.

The joint degree distribution $P(k, k')$ is the probability that an arbitrary edge connects a node with degree k to a node with degree k' [13]. The conditional probability $P(k|k')$, which is another way to express dependencies between node degrees, is the probability that an arbitrary neighbor of a node with degree k has a degree k' [13]

$$P(k|k') = \frac{\langle k \rangle \, P(k, k')}{kP(k)} \tag{2.9}$$

The average degree of the nearest neighbors of nodes with degree k can be computed as [13]

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) \tag{2.10}$$

If there are no correlations k_{nn} is independent of k, and such a network is called a neutral network [13]. If $k_{nn}(k)$ is an increasing function of k, then nodes with a high degree tend to connect to nodes with a high degree and the network is classified as assortative. If instead $k_{nn}(k)$ is a decreasing function of k, nodes with a high degree tend to connect to nodes with a low degree and the network is called disassortative [13].
However, the terms assortativity and disassortativity are also often used without taking degree correlations into account. Instead, one may look at a particular node property to define assortative and disassortative network [38]. In this way, an assortative network is a network where nodes tend to connect to other nodes with similar properties as themselves, whereas in a disassortative network nodes tend to connect to nodes with different properties [38].

2.1.5   Network distances

Several very useful measurements quantify the notion of length and distance in a network/graph. We start with the definition of a path: a path is a sequence of distinct nodes, such that adjacent nodes in the sequence are adjacent nodes in the network [19]. The length of a path is then defined as the number of edges in the path [19]. A geodesic path, often called the shortest path, between two nodes is the path between two nodes i and j with minimum length (note that the geodesic path is not necessarily unique) [13]. The length of the geodesic path between the two nodes i and j is called the geodesic distance or shortest distance d_{ij} [13]. Geodesic distance is often just called distance [19]. The notion of distance enables to define several network measurements such as average distance, diameter and radius.

Average distance

The average distance is the mean value of the geodesic distance d_{ij} [13]

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{2.11}$$

This definition, however, diverges if there are unconnected nodes in the network (unconnected nodes have, by definition, a distance equal to infinity) [13]. This problem could be avoided by only including connected pairs in the sum, but this introduces a distortion for networks with a lot of unconnected pairs of nodes. These networks will then have a small average distance, which is only expected for networks with lots of connections [13]. Instead, another definition of average distance is used: the global efficiency [13]

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{2.12}$$

This measurement quantifies the efficiency of the network in sending information between nodes [13]. This assumes that the efficiency of sending information between two nodes is reciprocal to their distance [13]. The harmonic mean of the geodesic distances is now defined as the reciprocal of the global efficiency

$$h = \frac{1}{E} \tag{2.13}$$

Diameter and radius

The diameter $diam(G)$ of a network G is the maximum distance between any pair of nodes [19]. The eccentricity of a node is the maximum distance from that node to any other node. The radius $rad(G)$ is the minimum eccentricity among all nodes of G. Note that the diameter is the maximum eccentricity among all nodes in G [19].

2.1.6 Clustering coefficient and transitivity

The clustering coefficient is another important measure of network topology. It is a measure for the number of triangles in a network and determines the connectivity in the neighborhood of a node i: if a node i has a high clustering coefficient, its neighbors are likely to be directly connected to each other [28].
Here a triangle is defined as a loop of length three, this is a sequence of nodes

x, y, z, x such that $\{x, y\}, \{y, z\}$ and $\{z, x\}$ are edges of the network. The clustering coefficient is thus a way to measure the degree to which nodes in a network tend to cluster [28]. The local clustering coefficient of a node i is defined as

$$C_i = \frac{2n_i}{k_i(k_i - 1)} \tag{2.14}$$

where n_i is the number of edges that actually exist between the nodes in the neighborhood of i and $k_i(k_i - 1)/2$ is the maximum number of edges that could exist between them (note that this expression is only valid for an undirected graph, since for a directed graph $e_{ij} \neq e_{ji}$ and we have $k_i(k_i-1)$ possible edges between the neighbors of node i). The average clustering coefficient is then given as the average of the local clustering coefficients

$$\bar{C} = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{2.15}$$

where N denotes, once again, the number of nodes in the network. The clustering coefficient is, beside a measure for the connectivity in the network, also linked to the robustness, or resilience against random damage, of the network [24] [28].

A measure that is closely related to the clustering coefficient is the transitivity T, defined as (valid for undirected, unweighted networks) [13]

$$T = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of nodes in the network}} \tag{2.16}$$

A connected triple is defined as a set of three nodes with at least two edges between them, so that each node can be reached from the other two (either directly or indirectly). The factor three arises from the fact that each triangle contributes to three different connected triples in the network: one centered at each node in the triangle [13]. Let us denote the number of triangles as N_Δ and the number of connected triples as N_3. These two numbers can be obtained from the adjacency matrix in the following way [13]

$$N_\Delta = \sum_{k>j>i} A_{ij} A_{ik} A_{jk} \tag{2.17}$$

$$N_3 = \sum_{k>j>i} \left(A_{ij} A_{ik} + A_{ji} A_{jk} + A_{ki} A_{kj} \right) \tag{2.18}$$

where A_{ij} are the elements of the adjacency matrix. It is possible to define the transitivity of one node as [13]

$$T_i = \frac{N_\Delta(i)}{N_3(i)} \tag{2.19}$$

where $N_\Delta(i)$ represents the number of triangles that involve node i and $N_3(i)$ is the number of connected triples with i as the central node [13]

$$N_\Delta(i) = \sum_{k>j} A_{ij} A_{ik} A_{jk} \qquad (2.20)$$

$$N_3(i) = \sum_{k>j} A_{ij} A_{ik} \qquad (2.21)$$

It is not too hard to see that $N_\Delta(i)$ counts the number of edges between the neighbors of i and that $N_3(i)$ is equal to $k_i(k_i - 1)/2$, where k_i is the degree of node i. It is thus obvious that the local clustering coefficient C_i (Eq. (2.14)) of a node and the transitivity T_i (Eq. (2.19)) of a node define the same quantity [13]. In the case of the average clustering coefficient $\bar{C}$ and the global transitivity T this is, however, not the case. The difference between the two definitions is that the average of Equation (2.16) gives the same weight to each triangle in the network, whereas Equation (2.15) gives the same weight to each node. This may lead to slightly different values since nodes with a higher degree may possibly be involved in a higher number of triangles than nodes with a lower degree [13].

2.1.7 Network communities

Network communities is another feature that (complex) networks may posses. A community inside a network is loosely defined as a set of nodes that are more densely connected to nodes inside that set than to the other nodes in the network [36]. More strictly, a community is defined based on two hypotheses: the connectedness hypothesis and the density hypothesis [4]. In short, the connectedness hypothesis means that each member of a community should be reached through each other member of the same community [4]. The density hypothesis implies that nodes inside a community are more likely to be linked to other nodes inside that community than to nodes outside the community [4]. The density hypothesis narrows what could be considered a community, but it doesn't uniquely define it. Several community definitions are consistent with the density hypothesis. Let us consider three possible definitions: maximum cliques, strong community and weak community [4].

Maximum cliques

A clique is a complete subgraph, where a complete subgraph is defined as a set of nodes of the network where each node in the set is directly connected to all the others nodes in the same set [4]. A community based on this definition would

then be the largest clique in the network. This definition of community might, however, be too restrictive and, beside that, large cliques don't appear very frequent in networks [4].

Strong communities

A strong community is defined such that each node inside the community has more edges to other nodes inside the same community than to nodes outside the community [4]. Let us denote a community as C and let i be a node inside the community C. If we define the internal degree $k_i^{int}(C)$ as the number of edges of node i with other nodes in C and the external degree $k_i^{ext}(C)$ as the number of edges of node i with nodes outside C, then we have the following condition for C to be a strong community [4]

$$k_i^{int}(C) > k_i^{ext}(C) \qquad \forall i \in C \tag{2.22}$$

Weak communities

A weak community is a community where the sum of the internal degree of all the nodes in the community exceeds the sum of the external degree of all the nodes in the community [4]. We thus have the following condition for a weak community C [4]

$$\sum_{i \in C} k_i^{int}(C) > \sum_{i \in C} k_i^{ext}(C) \tag{2.23}$$

2.1.8   Modularity

The modularity Q is a measurement that represents the quality of a particular division of a network into different communities [13]. It was proposed after the fundamental problem arised for real networks concerning how to best divide the network into its constituent communities [13]. Generally, no a priori information is available about the number of existing communities in real networks [13].
The modularity for a network with N nodes and L edges and partitioned into n_c communities is calculated as follows. Each community has N_c nodes and L_c edges ($c = 1, ..., n_c$) [4]. The modularity of a community Q_c is then calculated by measuring the difference between the network real wiring diagram (which is given by the adjacency matrix A_{ij}) and the expected number of edges between i and j if the network is randomly wired (where i and j are two nodes that lie in the same community C_c) [4]. This last quantity is given by p_{ij} which is obtained

by randomizing the original network while keeping the expected degree of each node unchanged [4]. Hence, the modularity of a community Q_c becomes [4]

$$Q_c = \frac{1}{2L} \sum_{(i,j) \in C_c} (A_{ij} - p_{ij}) \tag{2.24}$$

where the factor $\frac{1}{2L}$ is just a normalization factor. Using [4]

$$p_{ij} = \frac{k_i k_j}{2L} \tag{2.25}$$

we can derive a simpler form for Q_c [4]

$$Q_c = \frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \tag{2.26}$$

where L_c is the total number of edges in the community C_c and k_c is the total degree of the nodes in this community [4].

To generalize this idea to the whole network, we sum Eq. (2.26) over all n_c communities [4]

$$Q = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right] \tag{2.27}$$

The derivation of this formula and thus of Eq. (2.26) can be found in Appendix A.1, Section A.1.1. The modularity Q can have values in the interval $[-1, 1]$ and has the following properties [4]

- The higher the value of Q is for a partition, the better is the corresponding community structure. This idea is used in community detection algorithms, where modularity is maximized in order to find the optimal community structure.

- A modularity Q equal to zero corresponds to the situation where all nodes are put in the same community, this is if the whole network is put into a single community. Then the two terms in Eq. (2.27) become equal and Q vanishes.

- If every node is put into a separate community, the first term in Eq. (2.27) vanishes and we are left with n_c negative terms. Hence, this situation corresponds to a negative modularity.

As said before, modularity is often used in community detection algorithms. It is thus important to be aware of the limitations of modularity. One of the biggest

shortcomings of modularity maximization as community detection is its resolution limit. Modularity maximization forces small communities into larger ones [4]. If two communities A and B are merged, the modularity changes by [4]

$$\Delta Q_{AB} = \frac{l_{AB}}{L} - \frac{k_A k_B}{2L^2} \tag{2.28}$$

where l_{AB} is the number of direct edges that connect nodes in community A with nodes in community B; k_A and k_B are the total degree of nodes in communities A and B respectively. The derivation of this formula is found in Appendix A.1, Section A.1.2. If A and B are distinct communities, they should remain distinct when Q is maximized [4]. This is however not always the case. Consider the case where $\frac{k_A k_B}{2L} < 1$ and there is at least one edge between A and B ($l_{AB} \geqslant 1$), then $\Delta Q_{AB} \geqslant 0$ and the modularity increases by merging the two communities [4]. Assume for simplicity that $k_A = k_B = k$, then modularity is increased by merging A and B if

$$k \leqslant \sqrt{2L} \tag{2.29}$$

and if there is at least one edge between the two communities; even if A and B are distinct communities [4]! Equation (2.29) is called the resolution limit: communities that are smaller than this limit will not be detected by modularity maximization [4].

2.2   Network models

In order to study the topological structures of real-world networks, several theoretical network models have been developed. These theoretical models usually have a simpler representation and well known properties that can be derived analytically. They are widely studied and used extensively. The study of complex networks relies heavily on the knowledge and understanding of these models. In this section we will review some of the most important ones.

2.2.1   The Erdős-Rényi model

In 1959 the two mathematicians Paul Erdős and Alfréd Rényi proposed a model to generate simple random networks [13]. Independently of them, Solomonoff and Rapoport already proposed the model in 1951. The model constructs random networks with N nodes in the following way [2]

1. Start with N disconnected nodes,

2. For each pair of nodes, connect them with a predefined probability p.

If $p = 1$, we obtain a network with the maximum number of edges $N(N-1)/2$. This is called the fully connected network.

Degree distribution and average degree

The probability that a node is connected to k other nodes and not to $N - k$ others is given by

$$p^k(1-p)^{N-1-k} \tag{2.30}$$

The number of ways to choose the k nodes among the $N - 1$ possible candidates is given by the binomial coefficient

$$\binom{N-1}{k} = \frac{(N-1)!}{k!(N-1-k)!} \tag{2.31}$$

and the probability of being connected to exactly k other nodes becomes

$$p_k = \binom{N-1}{k} p^k(1-p)^{N-1-k} \tag{2.32}$$

We thus derived the degree distribution of a random network. From this we can calculate the average degree of the network. The expected mean degree is given by [41]

$$
\begin{aligned}
\langle k \rangle &= \sum_{k=0}^{N-1} k\, p_k \\
&= \sum_{k=0}^{N-1} k \binom{N-1}{k} p^k(1-p)^{N-1-k}
\end{aligned}
\tag{2.33}
$$

This equation can be simplified to $\langle k \rangle = (N-1)p$, and for $N \gg 1$ this approximately becomes $\langle k \rangle = Np$. The derivation of this result can be found in Appendix B.1, Section B.1.1.

Most real networks are sparse, which means that $\langle k \rangle \ll N$. In this case the degree distribution of the random network is well approximated by a Poisson distribution [3]

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{2.34}$$

The derivation can be found in Appendix B.1, Section B.1.2.

Average clustering coefficient

The average clustering coefficient (Eq. (2.16)) of the Erdős-Rényi network is given by [8]

$$\bar{C} = \frac{\binom{N}{3}p^3}{\binom{N}{3}p^2} = p = \frac{\langle k \rangle}{N - 1} \tag{2.35}$$

This can be easily understood since choosing three nodes out of N is given by $\binom{N}{3}$. The chance that they are connected by three edges is given by p^3, which gives the number of triangles; the chance that the are connected by two edges is p^2, which determines the number of connected triples. If the average degree $\langle k \rangle$ is constant, the average clustering constant goes as $\bar{C} \sim \mathcal{O}(1/N)$, which becomes very small for large networks [8].

Diameter and average distance

A heuristic derivation of the diameter of a random network goes as follows. Consider a random network with average degree $\langle k \rangle$, then a node in this network has on average [3]

- $\langle k \rangle$ nodes at distance one ($d = 1$)

- $\langle k \rangle^2$ nodes at distance two ($d = 2$)

- $\langle k \rangle^3$ nodes at distance three ($d = 3$)

- ...

- $\langle k \rangle^d$ nodes at distance d

The number of nodes at a distance d cannot exceed the number of nodes N in the network. We thus obtain the following equation from which the maximum distance or diameter d_{max} can be obtained

$$N \approx \langle k \rangle^{d_{max}} \tag{2.36}$$

or

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle} \tag{2.37}$$

A more precise calculation can be found in Appendix B.1, Section B.1.3. Equation (2.37) might, however, be a better approximation to the average distance $\langle d \rangle$ between two randomly chosen nodes [3]. This is because d_{max} is often dominated by a few extreme paths, where $\langle d \rangle$ instead is averaged over all node pairs and

thus suppresses these fluctuations [3]. Hence, the average distance of a random network is defined by [3]

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle} \tag{2.38}$$

Since for large N, $\ln N \ll N$, and we find that the average distance is a lot smaller than the size of the network. Generally, the logarithmic dependence of the average distance $\langle d \rangle$ on the network size N is one of the two defining properties of what is referred to as the "small world phenomenon"; the other property is a high clustering coefficient [16].

To summarize: the random Erdős-Rényi network has a bell shaped degree distribution which approximates a Poisson distribution for $N \gg 1$; the average clustering coefficient becomes small for large networks and the average distance depends logarithmic on the system sizes. The Erdős-Rényi network thus possesses one of the two properties of the "small world phenomenon" (namely short average distance).
Real social networks, however, usually display the "small world phenomenon" and thus have both a short average distance and a high clustering coefficient. The question then rises: can we construct a model that has both of these properties? The answer is yes and this is exactly what Duncan Watts and Steve Strogatz did.

2.2.2 The Watts-Strogatz model

In 1998, Duncan Watts and Steve Strogatz constructed a network model that generates random networks with small world properties, such as high clustering and short average distances. Their model achieves this by interpolating between a regular ring lattice and a random network [39]. The construction goes as follows [39]

1. Start with a regular ring lattice with N nodes, each connected to $K = 2m$ neighbors (m on each side)

2. For each node i, rewire each edge that connects i with its $K/2$ rightmost neighbors with a probability p. Rewiring is done in such a way that self-loops and edge duplication are avoided

The probability p allows us to tune the graph between a regular lattice ($p = 0$) and a random network ($p = 1$) [39]. Usually, the following condition is required for the number of neighbors K: $N \gg K \gg \ln N \gg 1$. The first inequality assures sparse networks, whereas the second inequality guarantees that a random graph will be connected [39].

The regular ring lattice ($p = 0$) has an average clustering coefficient $\bar{C}(0)$ and an average path length $\langle d(0) \rangle$ equal to [39]

$$\bar{C}(0) = \frac{3(K-2)}{4(K-1)} \approx \frac{3}{4} \text{ for } K \gg 1 \qquad \langle d(0) \rangle \approx \frac{N}{2K} \qquad (2.39)$$

The derivations of these formulas can be found in Appendix B.2, Section B.2.1. The regular ring lattice thus has a high clustering coefficient and an average path length that grows linearly with the system size N.

The random network ($p = 1$) on the other hand has a low clustering coefficient $\bar{C}(1)$ and a small average path length $\langle d(1) \rangle$ (see Eq. (2.35) and Eq. (2.38)) [39]

$$\bar{C}(1) \approx \frac{K}{N} \ll 1 \qquad\qquad \langle d(1) \rangle \approx \frac{\ln N}{\ln K} \qquad (2.40)$$

Watts and Strogatz then showed by numerical simulation that the average distance for the networks with $0 < p < 1$ is comparable with the average distance of a random graph, even for small p [31]. The clustering coefficient is approximately given by [5]

$$\bar{C} \approx \bar{C}(0)(1-p)^3 \qquad (2.41)$$

where $\bar{C}(0)$ is given by Equation (2.39). This can be seen in the following way

- The probability that none of the three edges of the original triangles in the regular lattice are rewired is $(1-p)^3$

- The probability that edges are rewired back to each other is negligible

For rather small values of p, the clustering coefficient remains high. The model of Watts and Strogatz is thus able to generate random networks with a high clustering coefficient and short average distances.

2.2.3   The stochastic block model

Another random graph model is one that produces networks containing a community structure, namely the stochastic block model (SBM). The stochastic block model originates in social sciences and was developed to describe group structures in friendship networks [18]. The idea behind the development of this model is that often in real networks, there are groups of nodes that are more densely interconnected with each other than with the rest of the network [13]. Hence, these networks posses a modular/community structure and this is exactly the structure that is easily modeled by the stochastic block model. The stochastic block model is also often used as a benchmark for community detection algorithms.

The stochastic block model in its most simple form is defined by the following parameters [9]

- The number of nodes N

- A partition of the node set $\{1, 2, ..., N\}$ into disjoint subsets $C_1, ..., C_k$; these groups are the communities

- A symmetric $k \times k$ matrix M; the diagonal elements M_{ii}, $i = 1, ..., k$ represent the edge probabilities inside the community C_i and the off-diagonal elements M_{ij}, $i, j = 1, ..., k$ with $i \neq j$, give the edge probabilities between nodes in community C_i and nodes in community C_j

There are no real restrictions to the edge probabilities, which makes the stochastic block model a very general model that is capable of reproducing many different network structures. However, the edge probabilities inside a community are often taken to be bigger than the probabilities between communities ($M_{ii} > M_{ij}$, $i \neq j$). In this way, the stochastic block model generates a network with a clear community structure. A stochastic block model that fulfills this condition is called an assortative SBM [27]. A disassortative stochastic block model refers to the case where the edge probabilities between the communities are higher than those inside the communities ($M_{ii} < M_{ij}$, $i \neq j$) [20].

The stochastic block model includes the Erdős-Rényi model [9]. This can be seen in the two different ways. First, consider the case where $k = 1$, thus we only have one group. The matrix M then reduces to one number p and we are left with the parameters N and p of the Erdős-Rényi model [9]. Second, assume we have more than one group or community, $k > 1$, and assume that every element of the matrix M is the same, $M_{ij} = p$ [9]. Again, we are left with the parameters N and p of the Erdős-Rényi model. In both cases, the mathematical properties of the Erdős-Rényi model are obtained: a Poisson degree distribution, an average distance that is logarithmic with the system size N, vanishing clustering coefficient for large N, etc [9]. The structure within the groups or communities of the stochastic block model is that of a random graph, while the structure between the communities is that of a random bipartite graph [9].

In this thesis, the stochastic block model is generated in the following way. First, the parameters of the model are specified. Then, for each community C_i an Erdős-Rényi network is constructed, where the number of nodes equals the size of the community C_i and the edge probability p is given by M_{ii}. Finally, edges are added between communities C_i and C_j with probability $p = M_{ij}$.

The stochastic block model can be fitted to data [9]. Given a choice of the number of communities k and an observed network G, the stochastic block model can be used to infer the latent community structure and the edge probability matrix M [9].

One of the shortcomings of fitting the SBM to infer the community structure, is that real-world networks often have a heterogeneous or power law degree distribution. Since the SBM is made of random networks inside each group and random bipartite networks between each pair of groups, the degree distribution of the full network is always a mixture of Poisson distributions [9]. The consequence of this is that if a network possesses a more skewed degree distribution, the SBM will put the nodes with a high degree in a small group, together with a high probability of connecting to other, larger groups [9]. The reason behind this is one of the main assumptions of the stochastic block model, namely the fact that the edge probabilities only depend on which group the node belongs to [9]. In other words, nodes that belong to the same group are stochastically equivalent, meaning that they have equivalent connectivity patterns to other nodes [9]. However, a network with a skewed degree distribution violates this assumption of edge independence and the SBM seeks to explain this over other kinds of latent group structures that we might care about [9]. Karrer and Newman proposed a variation of the SBM, the degree corrected SBM, that deals with this limitation [25].

2.3   Real-world social networks

The models discussed in the sections above, describe real-world networks in an idealized, simplified way. They may posses some of the characteristics of real-world networks, but they certainly do not always describe real-world networks appropriately. The question then arises: what are the main properties/characteristics of real-world networks, and more specifically of real-world social networks? Another important question that can be asked is whether or not there is a significant difference between off-line social networks and on-line social networks?

As already briefly discussed in the introduction, it is found that real-world networks often posses three characteristic properties. These are the two properties of the "small world phenomenon", namely a short average distance and a high clustering coefficient, and a heterogeneous or power law degree distribution [10]. These properties are valid for both off-line and on-line social networks [40]. Lets discuss them in a bit more detail.

Short average distance

In the late 1960's Stanley Milgram performed a, what is now famous, experiment. He asked randomly selected people in Boston and Omaha to forward a letter to a distant target person. The letter could, however, only be sent to acquaintances, thought to be closer to the target person [10]. The remarkable result of this experiment was that the average number of steps between the sender and the target person was only around six. This phenomenon is now often referred to as "six degrees of separation" [10].
This phenomenon is also observed in many other real-world networks: in most real-world networks it is possible to go from one node to another node through a number of edges that is small compared to the system size [10]. Such networks usually have an average distance that depends logarithmically on the system size instead of linearly. A short average distance facilitates a fast transmission of information [40].

High clustering coefficient

In real-world social networks, one often observes the pattern that two acquaintances/friends of an individual are also acquaintances/friends of each other [10]. In graph terminology, this is translated into a high clustering coefficient (which quantifies a large number of triangles in the network) [10]. This property is not only found in social networks, but also in a wide variety of different real-world networks [10].

Heterogeneous or power law degree distribution

Many real-world networks are found to have a power law degree distribution [10]

$$P(k) \sim k^{-\gamma} \tag{2.42}$$

where the exponent γ often lies between 2 and 3. Networks with a power law degree distribution are also called heterogeneous or scale free networks. The term scale free refers to the fact that there is no characteristic scale in the network. Compared to the homogeneous network, where each node has a degree close to the average degree $\langle k \rangle$, the heterogeneous network has many nodes with a low degree and a few nodes with a very high degree (these nodes are referred to as hubs). If the exponent γ is smaller or equal to two ($\gamma \leqslant 2$), the average degree diverges; if $2 < \gamma < 3$, there is a finite average degree, but a diverging variance [32].

As said earlier, these three main characteristics are the same for on- and off-line social networks. Another property that is found in most on- and off-line networks is the appearance of a giant component, this is, most nodes belong to the same connected component [26]. They also often posses a modular/community structure [17][29].

There are, however, some properties that are different between on-line and off-line networks as well. One of them might be the degree correlations/mixing pattern. It is commonly assumed and accepted that most real-life social networks have an assortative mixing pattern [22][40]. For a long time, it was thought that this is also valid for on-line social networks. However, recent research has shown that a lot of on-line social networks display disassortative mixture [22][40]. This discrepancy between real-life social networks and on-line social networks may be explained as follows. In real-life, 'ordinary people' often want to be friends with celebrities, however, the celebrities usually prefer to stay within their own circles [22][40]. The 'ordinary people' don't have access to these circles and we are left with the situation where 'ordinary people' interact with 'ordinary people' and celebrities hang out among other celebrities; hence, an assortative structure. In on-line social networks, 'ordinary people' can easily connect with the celebrity and vice-versa, the celebrity wants to show his/her importance/influence by the number of fans [22][40]. This leads to a disassortative mixing pattern.

Chapter 3

Opinion dynamics

The study of opinion dynamics in a group of people, as already mentioned in the introduction, has two major layers: a network layer and an opinion dynamics layer. The network layer, which describes the underlying structure and correlations between the agents/people, is intensively described in the previous chapter, Chapter 2. In this chapter we will focus on the second important layer, namely that of the opinion dynamics models. These models need to describe, in a simplified way, how opinions are formed and how they evolve in a group of people that interact in a certain way. Hence, these opinion dynamics are the dynamics that run on the underlying network structure. Several models have been proposed; here we will review some of the most important ones, with the main focus on the models used in this thesis. Some major attention will also be given to the content curation/filtering algorithms used by social media companies. This is because this thesis deals with opinion dynamics on social media, so the aspect of filtering cannot be ignored in the formation and evolution of opinions.

There are two major groups of opinion dynamics models, namely binary models and continuous approaches. These will be discussed in the following two sections. This thesis only deals with binary opinion dynamics models, so these will be thoroughly described; the continuous models will be quickly reviewed in order to give a more complete overview.

3.1   Binary models

Binary models are models where each agent can have one of the two opinions A or B (or, equivalently, 0 or 1). A binary model could reflect real-life cases where there exist two competing opinions, such as a vote in an election system with two competing candidates [33].

3.1.1   The majority model

One of the simplest opinion dynamics models is the binary majority model. In this model each agent has one of the two opinions A or B (or 0 or 1). Initially, these opinions can be equally distributed or not. If the distribution is not equal, but instead eg. 80/20, there is a predominant opinion and a minority opinion.
The following updating rule is then imposed on the network: each agent/individual adopts the majority opinion of his/her nearest neighbors [33]. Thus, initially each agent has an opinion, then the opinions are evolved in time: each individual iteratively updates his/her opinion according to the majority opinion of his/her direct neighbors [33].
Several studies have been performed on this kind of model, searching for answers on several questions such as: what are the conditions that drive such a system to converge to unanimity and does an opinion, that is, initially, a majority opinion, remain a majority opinion in the final state [33]? Most of these studies conclude that, at the end of the opinion evolution, a complete consensus state is obtained where all the agents adopt the same opinion [33]. This is even the case when the opinions are equally and randomly distributed in an Erdős-Rényi network (with the side note that the degree needs to be high enough), as was shown by Benjamini, Chan, O'Donnell, Tamuz and Tan [6]. It is also found that the rate of convergence increases with increasing edge density and that if the network is sparse enough, consensus will not be reached and the two opinions coexist [33].

3.2   Continuous models

3.3   Algorithmic personalization

On-line social platforms use algorithms to order and filter the posts that appear on the time line of an individual [35]. They do this in order to make the experience on their platforms more pleasant, interesting and convenient [35]. If social media companies didn't use any kind of filtering, there would be way to many content to handle and process for the social media user [7]. This in turn would make the on-line social experience a lot more unpleasant and inconvenient. The reason behind this is that people are cognitive and time constrained [35]; hence, there is only a limited amount of information that one can process [7].

Appendix A

Network measurements: Derivations

A.1   Modularity

A.1.1   Simple form of network modularity

Using [4]

$$Q_c = \frac{1}{2L} \sum_{(i,j) \in C_c} (A_{ij} - p_{ij}) \tag{A.1}$$

and [4]

$$p_{ij} = \frac{k_i k_j}{2L} \tag{A.2}$$

we can write the modularity of the whole network as [4]

$$Q = \frac{1}{2L} \sum_{i,j=1}^{N} \left(A_{ij} - \frac{k_i k_j}{2L} \right) \delta_{c_i, c_j} \tag{A.3}$$

where c_i represents the community to which node i belongs and where the Dirac-delta function δ_{c_i, c_j} expresses that only pairs of nodes that belong to the same community are taken into account. We can then rewrite the first term as a sum over communities [4]

$$\frac{1}{2L} \sum_{i,j=1}^{N} A_{ij} \delta_{c_i, c_j} = \sum_{c=1}^{n_c} \frac{1}{2L} \sum_{i,j \in C_c} A_{ij} = \sum_{c=1}^{n_c} \frac{L_c}{L} \tag{A.4}$$

where L_c is the number of edges in community C_c and the factor 2 disappears because each edge is counted twice in A_{ij}. Similarly, the second term in Eq. (A.3) can be rewritten as [4]

$$\frac{1}{2L} \sum_{i,j=1}^{N} \frac{k_i k_j}{2L} \delta_{c_i,c_j} = \sum_{c=1}^{n_c} \frac{1}{(2L)^2} \sum_{i,j \in C_c} k_i k_j = \sum_{c=1}^{n_c} \frac{k_c^2}{4L^2} \tag{A.5}$$

where k_c is the total degree of the nodes in community C_c. Combining Eq. (A.4) and Eq. (A.5) gives the final result [4]

$$Q = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right] \tag{A.6}$$

A.1.2 Modularity change after merging two communities

Consider two communities A and B and denote with k_A and k_B the total degree of communities A and B respectively [4]. The change in modularity after merging the two communities is calculated as follows (using Eq. (A.6)) [4]

$$\Delta Q_{AB} = \left[\frac{L_{AB}}{L} - \left(\frac{k_{AB}}{2L} \right)^2 \right] - \left[\frac{L_A}{L} - \left(\frac{k_A}{2L} \right)^2 + \frac{L_B}{2L} - \left(\frac{k_B}{2L} \right)^2 \right] \tag{A.7}$$

where $L_{AB} = L_A + L_B + l_{AB}$ is the total number of edges in the merged community; L_A and L_B are the number of edges in communities A and B respectively and l_{AB} is the number of direct edges between communities A and B [4]. $k_{AB} = k_A + k_B$ is the total degree of nodes in the merged community [4]. Inserting the formulas for L_{AB} and k_{AB} in Eq. (A.7) gives

$$\Delta Q_{AB} = \frac{l_{AB}}{L} - \frac{k_A k_B}{2L^2} \tag{A.8}$$

Appendix B

Network models: Derivations

B.1 The Erdős-Rényi model

B.1.1 Average degree

The expected mean degree is given by [41]

$$
\begin{aligned}
\langle k \rangle &= \sum_{k=0}^{N-1} k \, p_k \\
&= \sum_{k=0}^{N-1} k \binom{N-1}{k} p^k (1-p)^{N-1-k}
\end{aligned}
\tag{B.1}
$$

In order to simplify this equation we can make use of the following formula [41]

$$
(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}
\tag{B.2}
$$

Differentiating both sides of this equation gives us [41]

$$
\begin{aligned}
n(p+q)^{n-1} &= \sum_{k=0}^{n} \binom{n}{k} k p^{k-1} q^{n-k} \\
&= \frac{1}{p} \sum_{k=0}^{n} \binom{n}{k} k p^k q^{n-k}
\end{aligned}
\tag{B.3}
$$

Finally, by substituting $q = 1 - p$, we get [41]

$$n(p + (1-p))^{n-1} = \frac{1}{p}\sum_{k=0}^{n}\binom{n}{k}kp^k(1-p)^{n-k}$$

$$np = \sum_{k=0}^{n}\binom{n}{k}kp^k(1-p)^{n-k} \tag{B.4}$$

where the right hand side is equal to Equation (B.1) with $n = N - 1$ and we thus obtain $\langle k \rangle = (N-1)p$.

B.1.2   Poisson form of the degree distribution

We begin with the exact binomial distribution (Eq.(2.32))

$$p_k = \binom{N-1}{k}p^k(1-p)^{N-1-k} \tag{B.5}$$

which characterizes a random graph. In the case $k \ll N$, we can simplify the binomial in the following way [3]

$$\binom{N-1}{k} = \frac{(N-1)(N-1-1)...(N-1-k+1)}{k!} \approx \frac{(N-1)^k}{k!} \tag{B.6}$$

We also note the following equality [3]

$$\ln[(1-p)^{N-1-k}] = (N-1-k)\ln\left(1 - \frac{\langle k \rangle}{N-1}\right) \tag{B.7}$$

where we made use of the fact that $\langle k \rangle = (N-1)p$. Using the series expansion [3]

$$\ln(1+x) = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - ... \qquad \forall\, |x| \leqslant 1 \tag{B.8}$$

we get [3]

$$\ln[(1-p)^{N-1-k}] \approx (N-1-k)\frac{\langle k \rangle}{N-1} = -\langle k \rangle\left(1 - \frac{\langle k \rangle}{N-1}\right) \approx -\langle k \rangle \tag{B.9}$$

which is valid if $k \ll N$. We thus find [3]

$$(1-p)^{N-1-k} = e^{-\langle k \rangle} \tag{B.10}$$

Substituting Eq. (B.10) and Eq. (B.6) in Eq. (B.5) and using $\langle k \rangle = (N-1)p$, we finally obtain the Poisson form of the degree distribution [3]

$$
\begin{aligned}
p_k &= \binom{N-1}{k} p^k (1-p)^{N-1-k} = \frac{(N-1)^k}{k!} p^k \mathrm{e}^{-\langle k \rangle} \\
&= \frac{(N-1)^k}{k!} \left(\frac{\langle k \rangle}{N-1} \right)^k \mathrm{e}^{-\langle k \rangle} \\
&= \mathrm{e}^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}
\end{aligned}
\tag{B.11}
$$

B.1.3 Diameter

The expected number of nodes at a distance d from a starting node is [3]

$$
N(d) = 1 + \langle k \rangle + \langle k \rangle^2 + ... + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}
\tag{B.12}
$$

$N(d)$ cannot exceed the number of nodes N in the network. The maximum distance or diameter d_{max} of the network can thus be found by setting [3]

$$
N(d_{max}) \approx N
\tag{B.13}
$$

If $\langle k \rangle \gg 1$, we can neglect the -1 terms in Eq. (B.12), which gives [3]

$$
N \approx \langle k \rangle^{d_{max}}
\tag{B.14}
$$

and thus the diameter of a random network becomes [3]

$$
d_{max} \approx \frac{\ln N}{\ln \langle k \rangle}
\tag{B.15}
$$

B.2 The Watts-Strogatz model

B.2.1 The regular ring lattice

Clustering coefficient

A triangle in the regular ring lattice requires two edge traversals in the same direction and one in the opposite direction [34]. In the regular ring lattice every node has a fixed degree K, this means that each node is connected to $K/2$ neighbors on its left and $K/2$ neighbors on its right. It is then clear that the final, backwards step can span at most $K/2$ nodes [34]; otherwise we would end up at a

different node from which we started, which would definitely not form a triangle. Hence, the number of triangles for a given node is given by choosing the two forward target nodes from the $K/2$ possibilities [34]

$$\binom{\frac{K}{2}}{2} = \frac{1}{4}K\left(\frac{K}{2} - 1\right) \tag{B.16}$$

The number of connected triples per node is given by choosing two nodes out of the K neighbors [34]

$$\binom{K}{2} = \frac{1}{2}K(K - 1) \tag{B.17}$$

Since every node in the regular ring lattice is identical, we can easily find the total number of triangles and connected triples in the network by multiplying Eq. (B.16) and Eq. (B.17) with the number of nodes N. We then obtain the final result for the clustering coefficient in the regular ring lattice [34]

$$\bar{C} = \frac{3 \cdot N\frac{1}{4}K\left(\frac{K}{2} - 1\right)}{N\frac{1}{2}K(K - 1)} = \frac{3(K - 2)}{4(K - 1)} \tag{B.18}$$

Average distance

The general formula for the average distance is given by Eq. (2.11)

$$l = \frac{1}{N(N - 1)} \sum_{i \neq j} d_{ij} \tag{B.19}$$

where the sum runs over the shortest path between every pair of nodes. Assume we have a regular ring lattice with N nodes and each node has K neighbors. Now we can calculate the distance between node i and every other node in the ring lattice

- There are K nodes adjacent to node i, so there are K nodes with a distance $d_{ij} = 1$

- There are K nodes that are neighbors of the neighbors of node i, so there are K nodes with a distance $d_{ij} = 2$

- ...

- There are K nodes that have a distance $d_{ij} = \frac{N}{K}$ between them and node i; its easy to see that these are the nodes that are the furthest away from node i

So the average distance for node i becomes [1]

$$l_i = \frac{1}{N-1}K\left(1+2+3+...+\frac{N}{K}\right) \approx \frac{K}{N}\left(1+2+3+...+\frac{N}{K}\right) \quad \text{(B.20)}$$

where the last approximation holds for $N \gg 1$. Since every node in the regular ring lattice is identical we can easily find the average distance of the network [1]

$$l = \frac{1}{N}\sum_i l_i = \frac{1}{N}Nl_i \approx \frac{K}{N}\left(1+2+3+...+\frac{N}{K}\right) \quad \text{(B.21)}$$

Using [1]

$$1+2+3+...+x = \sum_{i=1}^{x} i = \frac{x(x+1)}{2} \approx \frac{x^2}{2} \quad \text{(B.22)}$$

where the last approximation again holds for $x \gg 1$; we finally obtain [1]

$$l \approx \frac{K}{N}\frac{\frac{N^2}{K^2}}{2} = \frac{N}{2K} \quad \text{(B.23)}$$

Bibliography

[1] Chapter 9: Graphs and networks.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, Jan. 2002.

[3] A.-L. Barabási. Network science: Random networks. Sept. 2014.

[4] A.-L. Barabási. *Network Science*, chapter 9, pages 320–378. Cambridge University Press, July 2016.

[5] A. Barrat and M. Weigt. On the properties of small-world network models. *Europ. Phys. J. B*, 547(13), 2000.

[6] I. Benjamini, S.-O. Chan, R. O'Donnell, O. Tamuz, and L.-Y. Tan. Convergence, unanimity and disagreement in majority dynamics on unimodular graphs and random graphs. *Stochastic Processes and their Applications, Elsevier BV*, 126(9):2719–2733, Sept. 2016.

[7] E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and Information Technology, Springer Science and Business Media LLC*, 15(3):209–227, June 2013.

[8] A. Clauset. Inference, models and simulation for complex systems: Lecture 13. Oct. 2011.

[9] A. Clauset. Network analysis and modeling: Lecture 6. 2017.

[10] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda. Complex systems: Analysis and models of real-world networks.

[11] L. da F. Costa. What is a complex network? (cdt-2). 2018.

[12] L. da F. Costa, O. N. O. Jr., G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiqueira, M. P. Viana, and L. E. C. da Rocha. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. 2007.

[13] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas. Characteriza-tion of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, Jan. 2007.

[14] A. S. da Mata. Complex networks: a mini-review. *Brazilian Journal of Physics*, 50(5):658–672, July 2020.

[15] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, June 2002.

[16] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, 2010.

[17] E. Ferrara. A large-scale community structure analysis in facebook. *EPJ Data Science*, 1(1), Nov. 2012.

[18] T. Funke and T. Becker. Stochastic block models: A comparison of variants and inference methods. *PLOS ONE*, 14(4), Apr. 2019.

[19] W. Goddard and O. R. Oellermann. Distance in graphs. In *Structural Analysis of Complex Networks*, pages 49–72. Birkhäuser Boston, Sept. 2010.

[20] D. Gribel, T. Vidal, and M. Gendreau. Assortative-constrained stochastic block models. Apr. 2020.

[21] P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, Oct. 2012.

[22] H.-B. Hu and X.-F. Wang. Disassortative mixing in online social networks. *EPL (Europhysics Letters)*, 86(1), Apr. 2009.

[23] A. Ioannis and T. Eleni. Statistical analysis of weighted networks. 2007.

[24] S. Iyer, T. Killingback, B. Sundaram, and Z. Wang. Attack robustness and centrality of complex networks. *PLOS ONE*, 8(4), Apr. 2013.

[25] B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E, American Physical Society (APS)*, 83(1), Jan. 2011.

[26] M. Latapy and C. Magnien. Measuring fundamental properties of real-world complex networks.

[27] C. Lee and D. J. Wilkinson. A review of stochastic block models and ex-tensions for graph clustering. *Applied Network Science, Springer Science and Business Media LLC*, 4(1), Dec. 2019.

[28] Y. Li, Y. Shang, and Y. Yang. Clustering coefficients of large networks. *Information Sciences*, 382-383:350–358, Mar. 2017.

[29] M. McGlohon, L. Akoglu, and C. Faloutsos. Statistical properties of social networks. In *Social Network Data Analytics*, pages 17–42. Springer US, 2011.

[30] L. Muchnik, S. Pei, L. C. Parra, S. D. S. Reis, J. S. A. Jr, S. Havlin, and H. A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific Reports*, 3(1), May 2013.

[31] M. E. J. Newman. Models of the small world: A review. *J. Stat. Phys.*, 101:819–841, 2000.

[32] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, Sept. 2005.

[33] V. X. Nguyen, G. Xiao, X.-J. Xu, Q. Wu, and C.-Y. Xia. Dynamics of opinion formation under majority rules on complex social networks. *Scientific Reports, Springer Science and Business Media LLC*, 10(1), Jan. 2020.

[34] K. Pelechrinis. Telcom2125: Network science and analysis, 2015.

[35] N. Perra and L. E. C. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, 9(1), May 2019.

[36] B. Saha, A. Mandal, S. B. Tripathy, and D. Mukherjee. Complex networks, communities and clustering: A survey. 2015.

[37] A. Sîrbu, V. Loreto, V. D. P. Servedio, and F. Tria. Opinion dynamics: models, extensions and external effects. 2016.

[38] G. Thedchanamoorthy, M. Piraveenan, D. Kasthuriratna, and U. Senanayake. Node assortativity in complex networks: An alternative approach. *Procedia Computer Science*, 29:2449–2461, 2014.

[39] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[40] D. Zhang and G. Guo. A comparison of online social networks and real-life social networks: A study of sina microblogging. *Mathematical Problems in Engineering*, pages 1–6, 2014.

[41] J. Zhang, D. Goldburt, and J. Hopcroft. Lecture 01: Large graphs. Jan. 2006.